IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

UTILITY PATENT APPLICATION FOR:

SELECTING NODES CLOSE TO ANOTHER NODE IN A NETWORK USING LOCATION INFORMATION FOR THE NODES

Inventors:

Zhichen Xu
1272 Glen Haven Drive
San Jose, CA 95129

Sujata Banerjee
1327 Elsona Drive
Sunnyvale, CA 94087

Sung-Ju Lee
2260 Homestead Ct. #212
Los Altos, CA 94024

HP Docket No. 200401879-1

SELECTING NODES CLOSE TO ANOTHER NODE IN A NETWORK USING LOCATION INFORMATION FOR THE NODES

TECHNICAL FIELD

This invention relates generally to networks. More particularly, the invention relates to using location information for nodes in a network.

BACKGROUND

The Internet, as it has grown considerably in size and popularity, is being used to provide various services and applications to users. Diverse applications, such as streaming a short movie demonstrating how to assemble a piece of furniture, taking a virtual tour of a real estate property or a scenic spot, watching a live performance of an artist, and participating in a networked multi-user computer game or conference, are all available to users via the Internet.

An important trend is that users are no longer satisfied with receiving services that are targeted at mass audiences. Users are demanding services that are tailored to their individual needs. With the proliferation of personalized services, an important challenge facing future network infrastructure is balancing the tradeoffs between providing individualized services to each user and making efficient use of network resources.

A fundamental challenge in effectively utilizing network resources and services is efficiently and quickly locating desired resources/services in large networks, such as the Internet. For example, a user may generate a query for finding particular media content available in the network. Location and distance estimation techniques may be used to find the closest cache or proxy in the network that provides the desired data or service, such as the desired media content.

Landmark clustering is a known location and distance estimation technique for determining a distance to a node in a network. Landmark clustering was introduced for routing in large networks. A node's physical location in a network is estimated by determining the node's distance to a common set of landmark nodes in the network. Landmark clustering assumes that if two nodes have similar distances (e.g., measured latencies) to the landmark nodes, the two nodes are likely to be close to each other. Routers store the estimated physical locations of the nodes and use the position information for routing to the closest node.

Figure 11 illustrates estimating physical position for the nodes 1110 and 1120 in the network 1100 using landmark clustering. The client nodes 1110 and 1120 determine their distances to the landmark nodes L1101 and L1102. Because the nodes 1110 and 1120 have similar distances to the landmark nodes L1101 and L1102, the nodes 1110 and 1120 are determined to be close to each other.

Landmark clustering is an efficient technique for determining general location information for nodes. However, current landmark clustering techniques tend to result in false clustering, where nodes that are far away in network distance are clustered near each other. That is nodes that are far away from landmark nodes tend be estimated as having locations near each other when in fact they are located at substantial distances from each other. Secondly, landmark clustering is a coarse-grained approximation and is not very effective in differentiating between nodes that are relatively close in distance.

SUMMARY OF THE EMBODIMENTS OF THE INVENTION

According to an embodiment, a method of identifying at least one node close to a first node in a network includes selecting a set of candidate nodes from a plurality of nodes based on location information for the candidate nodes and the first node. The method also includes applying a clustering algorithm to the location information for the candidate nodes and the first node, and identifying a subset of the set of candidate nodes closest to the first node based on the results of applying the clustering algorithm.

According to another embodiment, a node in a network includes means for selecting a set of candidate nodes from a plurality of nodes based on location information for the candidate nodes and a first node. The node also includes means for applying a clustering algorithm to the location information for the candidate nodes and the first node, and means for identifying a subset of the set of candidate nodes closest to the first node based on the results of applying the clustering algorithm.

According to yet another embodiment, a computer system is operable to connect to a peer-to-peer network and is operable to function as a distributed hash table node in a distributed hash table overlay network. The distributed hash table overlay network is a logical representation of the peer-to-peer network. The computer system includes amemory operable to store location information for a plurality of nodes in the peer-to-peer network that are physically close to the computer system. The computer system also includes a processor operable to compare the location information for the plurality of nodes to location information for a first node to identify a set of nodes from the plurality of nodes that are physically close to the first node in the peer-to-peer network.

According to yet another embodiment, a method of storing information in a node in a network, wherein the node is operable to function as a distributed hash table node in a distributed hash table overlay network, includes receiving location information for a plurality of nodes. The nodes are located physically close in the network. The method also includes storing the location information in a table, wherein the location information for the plurality of nodes comprises distances measured from each of the plurality of nodes to a plurality of global landmark nodes and to at least one local landmark node.

BRIEF DESCRIPTION OF THE DRAWINGS

Various features of the embodiments can be more fully appreciated, as the same become better understood with reference to the following detailed description of the embodiments when considered in connection with the accompanying figures, in which:

Figure 1 illustrates using global landmark nodes and local landmark nodes in a network to generate location information according to an embodiment;

Figure 2 illustrates a 2-dimensional CAN overlay network for the network shown in figure 1, according to an embodiment;

Figure 3 illustrates a landmark space including landmark vectors, according to an embodiment;

Figure 4 illustrates using a hash function to translate points in the landmark space shown in figure 3 to an overlay network, such as shown in figure 2, according to an embodiment;

Figure 5 illustrates another network, according to an embodiment;

Figure 6 illustrates an overlay network for the network shown in figure 5, according to an embodiment;

Figure 7 illustrates a flow chart of a method for determining location information for a node in a network, according to an embodiment;

Figure 8 illustrates a flow chart of a method for determining a closest node to a given node, according to an embodiment;

Figure 9 illustrates a peer-to-peer system, according to an embodiment;

Figure 10 illustrates a computer system that may operate as a node in the peer-to-peer system shown in figure 9, according to an embodiment; and

Figure 11 illustrates a conventional landmark clustering scheme.

DETAILED DESCRIPTION OF THE EMBODIMENTS

For simplicity and illustrative purposes, the principles of the embodiments are described. However, one of ordinary skill in the art would readily recognize that the same principles are equally applicable to, and can be implemented in, all types of network systems, and that any such variations do not depart from the true spirit and scope of the embodiments of the invention. Moreover, in the following detailed description, references are made to the accompanying figures, which illustrate specific embodiments. Electrical, mechanical, logical and structural changes may be made to the embodiments without departing from the spirit and scope of the embodiments of the invention.

According to an embodiment, an enhanced landmark clustering technique is used to estimate physical locations of nodes in a network. A node is any device that may send and/or receive messages from another device via the network. A physical location of a node, also referred to herein as the node's location in the network, is the node's location in the network relative to other nodes in the network. For example, location information for the node may be determined by measuring distances to other nodes in the network, such as global landmark nodes and local landmark nodes that are proximally located to the node. The location information may be used as an estimation of the node's physical location in the network. Distance to a node, for example, may be measured using a network metric such as round-trip-time or network hops. Distances between nodes and location information for nodes may not be the same as geographical distances between nodes and geographical locations of the nodes, because distances are measured in terms of a network metric, such as round-trip-time or network hops, and not measured in terms of a geographical distance metric, such as kilometers or miles.

Global landmark nodes and local landmark nodes may be randomly selected from the nodes in a network. Almost any node in the network may be selected to be a global landmark node or a local landmark node. The number of nodes selected to be local landmark nodes and global landmark nodes is generally much smaller than the total number of nodes in the network. Also, the total number of global landmark nodes in the network is generally smaller than the number of local landmark nodes. The number of global and local landmark nodes used in the network may depend on the desired accuracy of the location information. To minimize network traffic local landmark nodes may be strategically placed in the network, such as near gateway routers. For example, routers encountered by a message from the node en route to a global landmark can be used as local landmark nodes.

As described above, location information for a node may be determined by measuring distances to global landmark nodes and local landmark nodes that are proximally located to the node. In one embodiment, the node measures distances to each of the global landmark nodes and the proximally located local landmark nodes in the network to determine the node's location information. In one example, a proximally located local landmark mark node is generally closer to the node than at least one of the global landmark nodes. For example, a local landmark node may be on a routing path between the node and a global landmark node. In this example, the distance to the local landmark nodes can be obtained with little or no added messaging overhead if these local landmark nodes can respond to measurement traffic, such as a probe packet for measuring round-trip-time. That is additional distance measurement traffic to the local landmark nodes need not be generated, because this example may utilize a probe packet being transmitted to a global landmark node to measure distances to local landmark nodes encountered en route to the global landmark node. In another example, a local landmark node may be proximally located to a node if the local landmark node is within a predetermined distance to the node. In this example, a node may identify local landmark nodes in proximity to the node using a global information table, and then measure distances to the identified local landmark nodes. Thus, local landmark nodes, which may not be on a routing path to a global landmark node but which may still be useful for accurately determining location information for the node, can be used.

Location information may be generated for substantially all the nodes in a network. The location information may be used for a variety of applications. For example, the location information may be used to identify a node for routing in the network. In another example, the location information may be used to find a closet node providing desired content or services for a user.

Compared with conventional landmark clustering techniques, the landmark technique according to embodiments of the invention a physical location of a node can be accurately estimated by determining distances to a relatively small number of global and local landmark nodes. Also, the local landmark nodes provide accurate information of the local network characteristics. Thus, optimal paths for routing based on local network characteristics may be selected.

According to another embodiment, location information for a given node and other nodes in the network is used to select a closest to node to the given node in the network. Closeness of a given node to other nodes in a network is related to the physical locations of the nodes. As described above, physical locations may be estimated based on distances between the nodes.

In one example, a set of candidate nodes close to a given node are selected. The candidate nodes may be selected by comparing location information for the given node to the candidate nodes. After the candidate nodes are selected, a clustering algorithm is applied to the location information of the candidate nodes to select a subset of the candidate nodes that are close to the given node. The given node may measure distances to the subset of candidate nodes to identify a closest node to the given node. Thus, after applying the clustering algorithm, distances to a minimal number of nodes can be used to accurately identify a closest node while generating minimal network traffic caused by measuring distances to nodes.

Figure 1 illustrates an example of using global landmark nodes and local landmark nodes in a network to generate location information. Location information is generated for nodes 10 and 20 in the network 100 by measuring distance to global landmark nodes and local landmark nodes in proximity to the nodes 10 and 20. For example, for node 10 distances are measured to the global landmarks GL1 and GL2. Distances are also measured to the local landmark nodes LL1 and LL2. Distance to a node may be measured using a known network metric, such as round-trip-time (RTT) or network hops. For example, the node 10 may transmit a probe packet to the global landmark node GL1 and measure RTT of the probe packet to determine the distance to the global landmark node GL1. A probe packet, for example, is a packet generated by node to measure one or more predetermined network metrics, such as RTT.

A landmark vector representing the location information for the node 10 is generated including the distances to the global landmark nodes GL1 and GL2 and the local landmark nodes LL1 and LL4. The landmark vector for the node 10 may be represented as <d(n, GL1), d(n, LL1), d(n, GL2), d(n, LL4)>, where d is the distance between the nodes and n represents the node for which location information is being generated.

Similarly, location information may be generated for the node 20. For example, distances are measured to the global landmarks GL1 and GL2. Distances are also measured to the local landmark nodes LL2 and LL3. A landmark vector representing the location information for the node 20 is generated including the distances to the global landmark nodes GL1 and GL2 and the local landmark nodes LL2 and LL3. The landmark vector for the node 20 may be represented as <d(n, GL1), d(n, LL2), d(n, GL2), d(n, LL3)>.

A location estimation technique that only considers distance to the global landmarks GL1 and GL2 may conclude that nodes 10 and 20 are in close proximity in the network 100, because the nodes 10 and 20 have similar distances to the global landmark nodes GL1 and GL2. These types of inaccuracies are known as false clustering. By accounting for the distances to the local landmark nodes LL1-LL4, false clustering is minimized and a more accurate estimation of the location of the nodes 10 and 20 is determined.

The network 100 may include many local landmark nodes and global landmark nodes, not shown. The number of nodes selected to be local landmark nodes and global landmark nodes is generally much smaller than the total number of nodes in the network. Also, the total number of global landmark nodes in the network 100 is generally smaller than the number of local landmark nodes. The number of global and local landmark nodes used in the network 100 may depend on the desired accuracy of the location information. Simulations have shown that a relatively small number of global landmarks are needed, for example, 15 global landmark nodes for a network of 10,000 nodes, to generate accurate location information. Almost any node in the network 100 may be chosen to be a global landmark node or a local landmark node. For example, a predetermined number of nodes in the network may be randomly selected to be global landmark nodes and local landmark nodes, whereby the number of global landmark nodes is smaller than the number of local landmark nodes. To minimize network traffic local landmark nodes may be strategically placed in the network 100, such as near gateway routers. For example, nodes near gateway routers may be selected to be local landmark nodes.

As described above, the nodes 10 and 20 measure distance to local landmark nodes proximally located to the nodes 10 and 20. In one embodiment, local landmark nodes are proximally located to a node if the local landmark nodes are on a routing path to a global node. For example, node 10 transmits a probe packet to the global landmark node GL1. The probe packet encounters local landmark node LL1, because it is on the routing path R1 to the global landmark node GL1. The local landmark node LL1 transmits and acknowledge (ACK) message back to the node 10. The node 10 determines distance to the local landmark node LL1, for example, using the RTT of the probe packet and the ACK message. Also, to minimize network traffic, a probe packet may keep track of the number of local landmark nodes that it has encountered, for example, by updating a field in a packet header similar to a time-to-live field. If a local landmark node receives a probe packet that has already encountered a predetermined number of local landmark nodes, the local landmark node simply forwards the packet without transmitting an ACK message.

In another embodiment, each of the local landmark nodes measures its distance to global landmark nodes to obtain its own landmark vector. These landmark vectors are stored in a global information table that is stored in the nodes in the network 100. The global information table is queried to identify local landmark nodes in proximity to a node. For example, the node 10 queries the global information table to identify local landmark nodes, such as the local landmark nodes LL1 and LL4 in proximity with the node 10. This may include identifying local landmark nodes having landmark vectors with a predetermined similarity to the node 10, wherein the predetermined similarity is related to a distance threshold between the node and the landmark node. Then, the node 10 determines distance to the local landmark nodes LL1 and LL4. Thus, a local landmark node need not be in a routing path to a global landmark node to be considered proximally located to the node 10.

Each node in the network 100 may generate location information, such as landmark vectors, by determining distances to the global landmark nodes and proximally located local landmark nodes. Each node stores its location information in a global information table. Thus, the global information table may include landmark vectors for substantially all the nodes in the network.

According to an embodiment, the global information table is implemented using a distributed hash table (DHT) overlay network. DHT overlay networks are logical representations of an underlying physical network, such as the network 100, which provide, among other types of functionality, data placement, information retrieval, and routing. DHT overlay networks have several desirable properties, such as scalability, fault-tolerance, and low management cost. Some examples of DHT overlay networks that may be used in the embodiments of the invention include content-addressable-network (CAN), PASTRY, CHORD, and expressway routing CAN (eCAN), which is a hierarchical version of CAN. The eCAN overlay network is further described in U.S. Patent Application Serial Number 10/231,184, entitled, "Expressway Routing Among Peers", filed on August 29, 2002, having a common assignee as the present application, and is hereby incorporated by reference in its entirety.

A DHT overlay network provides a hash table abstraction that maps keys to values. For example, data is represented in an overlay network as a (key, value) pair, such as (K1,V1). K1 is deterministically mapped to a point P in the overlay network using a hash function, e.g., P = h(K1). An example of a hash function is checksum or a space filling curve when hashing to spaces of different dimensions. The key value pair (K1, V1) is then stored at the point P in the overlay network, i.e., at the node owning the zone where point P lies. The same hash function is used to retrieve data. The hash function is also used for retrieving data from the DHT overlay network. For example, the hash function is used to calculate the point P from K1. Then the data is retrieved from the point P.

In one example, the global information table is stored in a CAN overlay network, however other types of DHT overlay networks may be used. In this example, a landmark vector or a portion of the landmark vector for a node is used as a key to identify a location in the DHT overlay network for storing information about the node. By using the landmark vector as a key, information about nodes physically close to each other in the underlying physical network are stored close to each other in the DHT overlay network, resulting in a minimal amount of traffic being generated when identifying a set of nodes close to a given node in the network.

Figure 2 illustrates an example of a 2-dimensional CAN overlay network 200 shown in figure 2, which is a logical representation of the underlying physical network 100. The nodes 30-50 shown in figure 2 are not shown in the network 100 shown in figure 1, but the nodes 30-50 may also be in the network 100. A CAN overlay network logically represents the underlying physical network using a d-dimensional Cartesian coordinate space on a d-torus. Figure 2 illustrates a 2-dimensional [0,1] x [0,1] Cartesian coordinate space in the overlay network 200. The coordinates for the zones 210-214 are shown. The Cartesian space is partitioned into CAN zones 210-214 owned by nodes 10-50, respectively. Each DHT node in the overlay network owns a zone. The nodes 30 and 20 are neighbor nodes to the node 10 and the nodes 40-50 and 10 are neighbor nodes to the node 20. Two nodes are neighbors if their zones overlap along d-1 dimensions and abut along one dimension. For example, the zones 210 and 214 abut along [0, .5] x [.5, 0]. The zones 210 and 213 are not neighbor zones because these zones do not abut along a dimension.

The nodes 10-50 each maintain a coordinate routing table that may include the IP address and the zone coordinates in the overlay network of each of its immediate neighbors. The routing table is used for routing from a source node to a destination node through neighboring nodes in the DHT overlay network 200. Assume the node 20 is retrieving data from a point P in the zone 214 owned by the node 30. Because the point P is not in the zone 211or any of the neighboring zones of the zone 211, the request for data is routed through a neighboring zone, such as the zone 213 owned by the node 40 to the node 30 owning the zone 214 where point P lies to retrieve the data. Thus, a CAN message includes destination coordinates, such as the coordinates for the point P determined using the hash function, for routing.

The global information table includes information about the nodes in the network 100, and the information is stored in the nodes in the DHT overlay network 200. To store information about a node in the global information table, the landmark vector for the node, which includes distances to the global landmark nodes in the network and distances to proximally located local landmark nodes, is used as a key to identify a location in the DHT overlay network for storing information about the node. By using the landmark vector or a portion of the landmark vector, such as the distances to the global landmark nodes, as a key, information about nodes physically close to each other in the network are stored close to each other in the DHT overlay network. Figure 3 illustrates a landmark space 300 including landmark vectors 310 and 320 for the nodes 10 and 20. The landmark space 300 is a logical representation of a space for mapping the landmark vectors of the nodes in the network 100.

The landmark space 300 is being shown to illustrate the mapping of the landmark vectors to locations in the DHT overlay network 200 for storing information in the global information table.

The global landmark portions of the landmark vectors for the nodes 10 and 20 are used to identify points in the landmark space 300 that are mapped to the DHT overlay network 100 for storing information in the global information table. The global landmark portion for the nodes 10 is <d(n, GL1), d(n, GL2))>, where d is distance to the global landmark nodes and n is the node 10 or 20. Each node in the network 100 may be mapped to the landmark space using the global landmark portion of the respective landmark vector. Also, the landmark space 300 may be much greater than two dimensions. The number of dimensions may be equal to the number of global landmark nodes used in the network 100. The nodes 10 and 20 are positioned in the landmark space 300 at coordinates based on their landmark vectors. Thus, nodes close to each other in the landmark space 300 are close in the physical network 100.

A hash function is used to translate physical node location information (e.g., landmark vectors) from the landmark space 300 to the overlay network 200, such that points close in the landmark space 300 are mapped to points that are close in the DHT overlay network 200. Figure 4 illustrates using a hash function to translate the points for the nodes 10 and 20 in the landmark space 300 to the overlay network 200. The hash function is used to determine the points 10' and 20' in the overlay network 200 that correspond to the points in the landmark space 300 for the nodes 10 and 20. The information for the nodes 10 and 20 is stored in the nodes that own the zone where the points 10' and 20' are located. Thus, by hashing the global landmark portion of a landmark vector, a node in the overlay network 200 is identified for storing information in the global information table, such as the complete landmark vector and other information associated with the nodes. Thus, the global information table is stored among the nodes in the DHT overlay network 200. Using a DHT overlay network to store landmark vectors is further described in U. S. Patent Application Serial Number 10/666,621, entitled "Utilizing Proximity Information in an Overlay Network" by Tang et al., having a common assignee with the present application, which is hereby incorporated by reference in its entirety.

In certain instances, the number of dimensions of the landmark space may be larger than the number of dimensions of the overlay network. A hash function comprising a space filling curve may be used to map points from the larger dimension space to the smaller dimension space, which is also described in the aforementioned patent application, U. S. Patent Application Serial Number 10/666,621, incorporated by reference.

The location information for the nodes in the network 100, such as landmark vectors including distances to the global landmark nodes and proximally located local landmark nodes, may be used to identify a closest node for a given node. For example, referring to figure 1, node 10 may be a client node used by a user. The node 10 determines its landmark vector, including distances to the global landmark nodes and proximally located local landmark nodes. The node 10 submits a request to find a close node to the DHT overlay network 100. For example, the node 10 hashes the global portion of its landmark vector, such as the distances to the global landmark nodes GL1 and GL2, to identify a point in the DHT overlay network 200. The node 10 transmits the request to a node in the DHT overlay network 200 owning the zone where the identified point is located.

The DHT overlay network 100 selects a set of candidate nodes closest to the node 10. For example, assume the node 10 transmits the request to the node 50. The node 50 selects a set of candidate nodes closest to the node 10 using the global landmark portion of the landmark vector for the node 10. For example, the node 50 compares the global landmark portion of the landmark vector of the node 10 to the global landmark portions of the landmark vectors for the nodes stored in the global information table residing in the node 50. This may include landmark vectors for nodes in the zone 212 owned by the node 50 and neighbor nodes to the node 50.

One measure of similarity between the landmark vectors or the global portions of the landmark vectors is the cosine of the angle between two landmark vectors. Landmark vectors that are the most similar to the landmark vector of a node, for example the node 10, may be selected as a candidate node. The number of candidate nodes selected may be based on the desired accuracy for finding the closest node.

Using the complete landmark vectors of all the candidate nodes and the complete landmark vector of the node 10, the node 50 applies a clustering algorithm to identify a subset of the set of candidate nodes that are closest to the node 10. A clustering algorithm is any algorithm that may be used to identify a subset of values from an initial set of values based on predetermined characteristics, such as similarities between location information. Four examples of clustering algorithms, described below by way of example and not limitation, are min_sum, max_dif f, order, and inner product.

The min_sum clustering algorithm assumes that if there are a sufficient number of landmark nodes, global and local, that two nodes n and c measure distances against, it is likely one of the landmark nodes, L, is located on a shortest path between the two nodes n and c, where n is a given node, such as the node 10, and c is a node in the initial set of candidate nodes determined to be close to the node 10. An example of the node L is the global landmark node GL1 located on the shortest path between the nodes 10 and 20.

For min_sum, the sum of dist(n, L) and dist(c, L) should be minimal. For the node n and its initial set of candidate nodes, represented as C, min_sum (n, C) is formally defined using equation 1 as follows:

Equation (1) $\min_{c \in C : L \in L(n,c)} (\mathrm{dist}(n, L) + \mathrm{dist}(c, L))$.

In equation 1, C is the set of candidate nodes, c is an element of the set C, and L(n, c) is the common set of landmark nodes, global and local, that the nodes n and c have measured against. Using equation 1, nodes from the candidate nodes C are selected for the subset of top candidates closest to the node n if they have the smallest distance sums for dist(n, L) + dist(c, L). Similarly, the assumption behind max_dif f is that if there are sufficient number of landmark nodes, global and local, that both n and c measure distances against, then there is a large likelihood that there exists a landmark node L such that c is on the shortest path from n to L or n is on the shortest path between c and L. In that case the ABS(dist(n, L) -dist(c, L)) may be used to identify a subset of the candidate nodes closest to the node n. The function ABS(x) returns the absolute value of x. Max_dif f(n, C) is formally defined using equation 2 as follows:

Equation (2) $\max_{c \in C : L \in L(n,c)} \mathrm{ABS}(\mathrm{dist}(n, L) - \mathrm{dist}(c, L))$.

For order, which is another example of a clustering algorithm, an assumption is made that if two nodes have similar distances to a set of common nodes, then the two nodes are likely to be close to each other. Using the order clustering algorithm, a node measures its RTT to the global landmark nodes and sorts the global landmark nodes in increasing RTTs. Therefore, each node has an associated order of global landmark nodes. Nodes with the same or similar order of global landmark nodes are considered to be close to each other. This technique however, cannot differentiate between nodes with the same global landmark orders, and thus is prone to false clustering.

For the nodes n, c, and L, where L is an element of the set of landmark nodes, global or local, that is common to the landmark vectors of nodes n and c, represented as $L \in L(n, c)$, the order of global landmarks in the landmark vector for the node n is defined as the order of global landmark nodes in the sorted list of all nodes L(n, c) based on their distances to the node n. The order of global landmark nodes is similarly defined. Thus, the order(n, c) is defined in equation 3 as follows:

Equation (3) $\min \sum_{L \in L(n,c)} ABS(order(L)n - order(L)c)$.

The clustering algorithm inner_product assumes that if a landmark node is close to a node n, then that landmark node can give a better indication of the location of the node n in the network. For example, the landmark vector for the node 10 may be represented as <d(n, GL1), d(n, LL1), d(n, GL2), d(n, LL4)>, where d is the distance between the nodes and n represents the node 10. If d(n, LL1) is shorter than d(n, LL4), then d(n, LL1) is given more weight by the inner_product clustering algorithm when comparing landmark vectors for the node 10 and the landmark vectors for the candidate nodes. The inner_product (n, c) is defined in equation 4 as follows:

Equation (4) $\max \sum_{L \in L(n,c)} ((1.0/(\mathrm{dist}(n, L)^2)) \times ((1.0/(\mathrm{dist}(c, L)^2)))$.

The landmark clustering algorithms described above are examples of algorithms that may be used to identify a subset of the initial set of candidate nodes that are closest to a node n, such as the node 10 shown in figure 1. Other known clustering algorithms, such as k-means, principal component analysis, and latent semantic indexing, may be used to select the subset of candidate nodes.

After the subset of candidate nodes are identified, the subset of the candidate nodes, such as a list of the subset of nodes which may include the landmark vectors of the nodes in the subset, are transmitted to the node 10. The node 10 measures distances to each of the nodes in the subset and selects the node closest to the node 10. By using this technique, a node desiring to identify a closest node measures distances to a small number of landmark nodes. Also, the local landmark nodes provide accurate information of local network characteristics, which may be used to select optimal routing paths. In addition, after utilizing a clustering algorithm, a node measures distances to a small number of top candidate nodes identified using the clustering algorithm.

One application for selecting the closest node to a given node is for identifying desired data or services provided by a node in a network. This is illustrated with respect to figure 5. Figure 5 illustrates a network 500 including nodes 501-506. The network 500 is similar to the network 100 shown in figure 1 and may be represented as a DHT overlay network that stores the global information table. The node 506 provides content to users via the network 500. For example, the node 506 provides content in English. The nodes 503 and 504 may be service nodes that convert the content from the node 506 to German and make the content available to users in German. The service nodes are nodes that provide services to users. In this case, the service is providing content from the node 506 in German.

The node 501, for example, is used by a user that desires to view the content from the node 506 in German. The node 501 submits a request for the service to the DHT overlay network along with its landmark vector. The landmark vector is the location information for the node 501 and includes distances to global landmark nodes and local landmark nodes. Submitting the request to the overlay network, for example, includes the node 501 transmitting the request to the node 502. The node 502 may be the owner of the zone where the hashed landmark vector of the node 501 lies, and thus the request is transmitted to the node 502. The node 502 searches the global information table to identify a set of candidate nodes that meet the request. The request, in addition to including the type of service desired, may also include other measured or determined characteristics related to quality of service (QoS), such as bandwidth of a path connecting two nodes, current load of a node, and forward capacity of a node. In this example, the global information table, in addition to including landmark vectors for nodes in the network 500, stores information including services provided by nodes and QoS characteristics. It will be apparent to one of ordinary skill in the art that the global information table may be populated with other types of data that may aid in providing users with desired services.

The node 502 selects a candidate set of nodes that meet the request. The node 502 applies a clustering algorithm to the candidate set of nodes to identify a subset of nodes that are closest to the node 501. The node 502 transmits a list of the subset of candidate nodes, such as the nodes 503 and 504, providing the content in German. The node 501 measures distance to each of the nodes 503 and 504. The node 501 may also measure for desired QoS characteristics, such as bandwidth, current load of a node, and forward capacity of a node. Measurements may be performed concurrently. The node 501 selects a node from the subset of candidate nodes that best meets its needs, such as the node 503. This may include the closest node or the closest node that provides the best quality of service.

The networks 100 and 500 shown in figures 1 and 5 are examples provided to illustrate the principles of the invention. It will be apparent to one of ordinary skill in the art that the embodiments of the invention may be practiced in much larger networks, such as tens of thousands of nodes. Furthermore, to improve the accuracy of the location information, a greater number of global and local landmark nodes may be used for determining location information for nodes.

Figure 6 illustrates an example of using location information including distances measured to global landmark nodes and local landmark nodes to identify a closest node for routing. Figure 6 illustrates an example of a DHT overlay network 600 for the network 500 shown in figure 5. The DHT overlay network 600 is illustrated as a CAN overlay network, such as described with respect to figure 2, including the nodes 501-506 in regions 610-615, however other types of DHT overlay networks may be used. Location information for the nodes 501-506 is stored in the global information table in the nodes 501-506 in the overlay network 600, such as described with respect to figures 3 and 4. The location information may include landmark vectors including distances to global landmark nodes and local landmark nodes.

The node 501 transmits a message to the node 503 for requesting personalized services, such as the content in German, from the node 503. Because the node 503 is not in a neighboring zone to the zone 610 where the node 501 is located, the node 501 identifies a closest node in a neighboring zone, such as the region 611 for routing to the node 503. To identify a closest node in the zone 611, the node 503 transmits a request to a node in the region 611. The request includes location information, such as a landmark vector, for the node 501. The routing table for the node 501 may include at least one node in the region 611 for transmitting the request. The node in the region 611 receiving the request searches its global information table for a set of candidate nodes in the region 611 that are closest to the node 501 and applies a clustering algorithm to identify a subset of the candidate nodes closest to the node 501. The subset of nodes is transmitted to the node 501.

The node 501 determines distance to each of the subset of candidate nodes and selects a closest node, such as the node 502, for routing. The node 501 may consider factors other than physical location when identifying a node for routing. For example, the node 501 may also evaluate QoS characteristics, which may be stored in the global information table, for the set of candidate nodes. The node 501 may select a closest node having the desired QoS characteristics from the set of candidate nodes. Assuming the node 502 is selected, the node 502 forwards the request for service from the node 501 to the node 503 providing the service.

Figure 7 illustrates a flow chart of a method for generating location information for nodes, according to an embodiment. Figure 7 is described with respect to the network 100 shown in figure 1 and the overlay network 200 shown in figure 2 by way of example and not limitation. At step 701, the node 10 determines distances to the global landmark nodes in the network 100. For example, the node 10 measures distances to the global landmark nodes GL1 and GL2 using RTT or another network metric.

At step 702, the node 10 determines distances to local landmark nodes in proximity to the node 10. This may include the local landmark nodes LL1 and LL4 encountered by a probe packet measuring RTTs to the global landmark nodes GL1 and GL2. In another example, distances to all local landmark nodes within a predetermined distance to the node are determined using the global information table. This may be determined by comparing landmark vectors for nodes. Nodes with landmark vectors having a predetermined similarity are selected from the global information table.

Steps 701 and 702 may be performed together. For example, when the local landmark nodes reside on the routing path, probing the global landmark node gets the distances to the corresponding local landmarks with substantially no messaging overhead. For example, substantially all the routers in the network may be selected as local landmark nodes and traceroute or another similar network utility is used to obtain the distances to global and local landmark nodes on the routing path to the global landmark node. In this example, distance to every router on the routing path may not be measured. For example, a time-to-live field may be utilized, such that distances to only the first predetermined number of routers receiving the probe packet are measured. Alternatively, distances to, for example, the 1st, 2nd, 4th, 8th, and 16th routers are measured. Thus, distances to a number of routers less than the total number of routers on a routing path to a global landmark node may be measured.

At step 703, location information for the node 10 is generated using the distances to the global landmark nodes and the local landmark nodes. For example, a landmark vector is generated for the node 10 including the distances to the global landmark nodes GL1 and GL2 and the distances to the local landmark nodes LL1 and LL4.

At step 704, the node 10 stores its location information, such as its landmark vector, in the global information table. In one example, this includes hashing the global landmark portion of the landmark vector to identify a location in the DHT overlay network 200, shown in figure 2, for storing the location information and possibly other information about the node 10, such as services provided, load, forwarding capacity, etc., in the global information table.

Figure 8 illustrates a flow chart of method 800 for identifying a closest node according to an embodiment. Figure 8 is described with respect to the network 100 shown in figure 1 and the overlay network 200 shown in figure 2 by way of example and not limitation. At step 801, location information, such as a landmark vector including distances to the global landmark nodes and the proximally located local landmark nodes, is determined for the node 10.

At step 802, the node 10 transmits a request to the DHT overlay network 200 for identifying a closest node to the node 10. For example, the node 10 hashes the global portion of its landmark vector, such as the distances to the global landmark nodes GL1 and GL2, to identify a point in the DHT overlay network 200. The node 10 transmits the request to a node in the DHT overlay network 200 owning the zone where the identified point is located.

At step 803, the DHT overlay network 100 selects a set of candidate nodes closest to the node 10. For example, assume the node 10 transmits the request to the node 50. The node 50 selects a set of candidate nodes closest to the node 10 using the global landmark portion of the landmark vector for the node 10. For example, the node 50 compares the global landmark portion of the landmark vector of the node 10 to the global landmark portions of the landmark vectors for the nodes stored in the global information table residing in the node 50. This may include landmark vectors for nodes in the zone 212 owned by the node 50 and neighbor nodes to the node 50. The landmark vectors or the global portions of the landmark vectors are compared to identify a set of candidate nodes closest to the node 10.

At step 804, the node 50 applies a clustering algorithm to set of candidate nodes to identify a subset of the candidate nodes that are closest to the node10. For example, the landmark vector of the node 10 is compared to the landmark vectors of the candidate nodes using, for example, one of the clustering algorithms min_sum, max_dif f, order, and inner product.

At step 805, the node 10 receives a list of the subset of candidate nodes from the node 50 determined at the step 804 and determine distances to each of the nodes in the subset of candidate nodes using a network metric, such as RTT or network hops. At step 806, the node 10 selects the closest node from the subset of candidate nodes based on the measured distances.

Factors other than distance may be considered when selecting a network node. For example, if the node 10 is attempting to identify a node in the network 100 that is providing particular data or services, factors related to QoS that can affect the transmission of the data or the delivery of services may be considered. For example, bandwidth, current of load of the node supplying the data or services, and forward capacity of a node may be determined for each of the nodes in the set of candidate nodes.

The node 10 selects a node from the candidate set that best meets its needs, such as a closest node that provides the best QoS. Similarly, if the node 10 is identifying a closest node for routing to, for example, a region in an overlay network, bandwidth, forwarding capacity and other factors associated with network routing may be considered when selecting a node from the candidate set of nodes.

Figure 9 illustrates a peer-to-peer (P2P) communications model that may be used by the underlying physical network, such as the networks 100 and 500 shown in figures 1 and 5, according to an embodiment of the invention. P2P networks are commonly used as the underlying physical network for DHT overlay networks, such as the CAN DHT overlay networks 200 and 600 shown in figures 2 and 6 respectively. A P2P network 900 includes a plurality of nodes 910a...910n functioning as peers in a P2P system. The nodes 910a...910n exchange information among themselves and with other network nodes over a network 920. The nodes 910a...910n may also determine which nodes 910a...910n perform other functions of a peer in a P2P system, such as object search and retrieval, object placement, storing and maintaining the global information table, etc. Objects may include files, URLs, etc. The nodes 910a...910n may be computer systems (e.g., personal digital assistants, laptop computers, workstations, servers, and other similar devices) that have a network interface. The nodes 910a...010n may be further operable to execute one or more software applications (not shown) that include the capability to share information (e.g., data, applications, etc.) in a P2P manner and the capability to operate as nodes in a DHT overlay network.

The network 920 may be operable to provide a communication channel among the nodes 910a...910n. The network 920 may be implemented as a local area network, wide area network or combination thereof. The network 920 may implement wired protocols, such as Ethernet, token ring, etc., wireless protocols, such as Cellular Digital Packet Data, Mobitex, IEEE 802.11b, Bluetooth, Wireless Application Protocol, Global System for Mobiles, etc., or combination thereof.

Some of the information that may be stored in the nodes 910a...n is shown for node 910a. The node 910a stores a routing table 931, the global information table 932, and possibly measured QoS characteristics.

Figure 10 illustrates an exemplary block diagram of a computer system 1000 that may be used as a node in the P2P network 900 shown in figure 9. The computer system 1000 includes one or more processors, such as processor 1002, providing an execution platform for executing software.

Commands and data from the processor 1002 are communicated over a communication bus 1004. The computer system 1000 also includes a main memory 1006, such as a Random Access Memory (RAM), where software may be executed during runtime, and a secondary memory 1008. The secondary memory 1008 includes, for example, a hard disk drive 1010 and/or a removable storage drive 1012, representing a floppy diskette drive, a magnetic tape drive, a compact disk drive, etc., or a nonvolatile memory where a copy of the software may be stored. The secondary memory 1008 may also include ROM (read only memory), EPROM (erasable, programmable ROM), EEPROM (electrically erasable, programmable ROM). In addition to software, routing tables, the global information table, and measured QoS characteristics may be stored in the main memory 1006 and/or the secondary memory 1008. The removable storage drive 1012 reads from and/or writes to a removable storage unit 1014 in a well-known manner.

A user interfaces with the computer system 1000 with one or more input devices 108, such as a keyboard, a mouse, a stylus, and the like. The display adaptor 1022 interfaces with the communication bus 1004 and the display 1020 and receives display data from the processor 1002 and converts the display data into display commands for the display 1020. A network interface 1030 is provided for communicating with other nodes via the network 920 shown in figure 9. Also, sensors 1032 are provided for measuring QoS characteristics for the node, which may include forward capacity, load, bandwidth, etc.

One or more of the steps of the methods 700 and 800 may be implemented as software embedded on a computer readable medium, such as the memory 1006 and/or 1008, and executed on the computer system 1000. The steps may be embodied by a computer program, which may exist in a variety of forms both active and inactive. For example, they may exist as software program(s) comprised of program instructions in source code, object code, executable code or other formats for performing some of the steps. Any of the above may be embodied on a computer readable medium, which include storage devices and signals, in compressed or uncompressed form.

Examples of suitable computer readable storage devices include conventional computer system RAM (random access memory), ROM (read only memory), EPROM (erasable, programmable ROM), EEPROM (electrically erasable, programmable ROM), and magnetic or optical disks or tapes. Examples of computer readable signals, whether modulated using a carrier or not, are signals that a computer system hosting or running the computer program may be configured to access, including signals downloaded through the Internet or other networks. Concrete examples of the foregoing include distribution of the programs on a CD ROM or via Internet download. In a sense, the Internet itself, as an abstract entity, is a computer readable medium. The same is true of computer networks in general. It is therefore to be understood that those functions enumerated below may be performed by any electronic device capable of executing the above-described functions.

Some example of the steps that may be performed by the software may include steps for determining distances to generate location information. For example, the software instructs the processor 1002 to use other hardware for generating probe packets for measuring RTT to global landmark nodes to determine distance. In another example, the software may generate a request to the global information table for identifying local landmark nodes within a predetermined proximity and measure distances to those local landmark nodes. The software includes instructions for implementing the DHT overlay network and for storing information to the global information table in the DHT overlay network by hashing a landmark vector. Also, some or all of the steps of the method 800 for identifying a closest node may also be performed using software in the computer system 1000.

It will be readily apparent to one of ordinary skill in the art that other steps described herein may be performed by the software. For example, if the computer system 1000 is selected as a local landmark node, the computer system 1000 may respond to received probe packets by generating an ACK message transmitted back to a node. Thus, the node transmitting the probe packet is able to determine distances to proximally located landmark nodes.

While the invention has been described with reference to the exemplary embodiments thereof, those skilled in the art will be able to make various modifications to the described embodiments without departing from the true spirit and scope. For example, it will be apparent to one of ordinary skill in the art that the advantages of storing location information as described herein can be applied to many applications, such as information storage, load balancing, congestion control, meeting quality of service (QoS) guarantee, taking advantage of heterogeneity in storage capacity and forwarding capacity, etc. The terms and descriptions used herein are set forth by way of illustration only and are not meant as limitations. In particular, although the method has been described by examples, the steps of the method may be performed in a different order than illustrated or simultaneously. Those skilled in the art will recognize that these and other variations are possible within the spirit and scope as defined in the following claims and their equivalents.